Conservative Confidence Bounds in Safety, from Generalised Claims of Improvement & Statistical Evidence

Kizito Salako, Lorenzo Strigini
Centre for Software Reliability
City, University of London
Northampton Square EC1V 0HB, U.K.
{k.o.salako,l.strigini}@city.ac.uk

Xingyu Zhao
Department of Computer Science
University of Liverpool
Ashton Street L69 3BX, U.K.
xingyu.zhao@liverpool.ac.uk

***Abstract*—"Proven-in-use", "globally-at-least-equivalent", "stress-tested", are concepts that come up in diverse contexts in acceptance, certification or licensing of critical systems. Their common feature is that dependability claims for a system in a certain operational environment are supported, in part, by evidence – *viz* of successful operation – concerning different, though related, system[s] and/or environment[s], together with an auxiliary argument that the target system/environment offers the same, or improved, safety. We propose a formal probabilistic (Bayesian) organisation for these arguments. Through specific examples of evidence for the "improvement" argument above, we demonstrate scenarios in which formalising such arguments substantially increases confidence in the target system, and show why this is not always the case. Example scenarios concern vehicles and nuclear plants. Besides supporting stronger claims, the mathematical formalisation imposes precise statements of the bases for "improvement" claims: seemingly similar forms of prior beliefs are sometimes revealed to imply substantial differences in the claims they can support.**

***Index Terms*—Reliability claims, statistical testing, safety-critical systems, ultra-high reliability, conservative Bayesian inference, field testing, not worse than existing systems, software re-use, globally at least equivalent, proven in use.**

I. Introduction

In dependability assessment, it often happens that favourable evidence is available in the form of experience of dependable operation. However, this evidence might not exactly match the situation for which the assessment is sought. For instance, "proven in use" evidence plays an important, accepted role [1,2] in assessing many systems. But there is concern whether this evidence (of past use) is relevant to the claim made. Due to this concern, standard IEC61508 [1], for example, sets strict conditions for accepting such experience as valid evidence: it must concern an identical system, under identical conditions of use. Good behaviour of a slightly different system version, or in slightly different conditions, is not admitted as evidence. One might object, not unreasonably, that this is too Draconian. A small change in the system, or in its mode of use, does void claims that the previous experience is a sample of the same stochastic process that the dependability assessment tries to predict. Yet it is still relevant evidence. True, even small changes may radically reduce reliability; but this is rare. The evidence is still relevant, but a little less so; what is hard is quantifying the effect of this reduced relevance. This neglect of useful evidence is most disturbing in cases of "ultra-high reliability" [3–5], where evidence of safe/correct operation is routinely insufficient.

We noted in previous work [6,7] that a special case of interest is that in which there is evidence that the change has been *for the better*. A general scenario is: dependability (e.g., safety) claims are to be supported for a situation (i.e., a system and an environment it operates in), say B, based on statistical evidence of good operation in B, and of good operation in another situation A. We focus on the common cases in which what changed between A and B is the system and/or its environment of use. But our mathematical results apply to any case in which a claim of *confidence in improvement* (CII) – from A to B – is justified.

More precisely, we define a CII as confidence in a claim of B being "no worse than" A, rather than "strictly better". Thus defined, CII includes "proven in use" (PIU in what follows) arguments: these commonly only claim similar dependability in the target environment to that experienced in the environment of past use.

The above abstract scenario generalises the case of PIU arguments, to include other common cases where CII plays a role: e.g., 1) the case of stress testing (in the lab or in the field) being claimed to be relevant evidence for reliability assessment; or 2) analysis-based arguments that a system is "globally at least equivalent" (GALE) to a previous one [8]; or 3) general claims that the system in B is an improvement on that in A.

Extending our previous work cited [6,7], in this paper we focus on the crucial passage of translating informal beliefs in "B being better than A" into formal statements that faithfully represent the evidence supporting those beliefs. We show that different formal statements may sometimes produce substantial

This work was partly supported by the Intel Collaborative Research Institute on Safe Automated Vehicles (ICRI-SAVe), and UK DSTL through the project "Safety Argument for Learning-enabled Autonomous Underwater Vehicles".

differences in the claims supported for B. These differences might well be missed in informal safety arguments. To this aim, we propose new specific example scenarios of evidence supporting CII (Sec. IV), propose two mathematical formulations of CII applicable to these scenarios ("PK" statements, Sec. V) and demonstrate their implications on the claims that can be supported. Our contribution includes both these examples (useful for practitioners) and the insights that they bring about this approach to using CII-based arguments.

We study these new scenarios in the context of the method and assumptions of our previous papers [6,7]:

1) applying *conservative Bayesian inference* (CBI) [9–11]: a use of Bayesian inference that aims to avoid the risk of unwittingly over-optimistic assessment. While Bayesian inference requires its user to specify a full "prior distribution", CBI does not. Instead, it uses, as its input, limited constraints on the prior, which are easier for experts to argue on the basis of the evidence. This improves trust that the dependability claims are the result of the actual evidence, rather than artefacts of assumptions made for mathematical convenience.

2) treating the common situation in which the claim of interest concerns a *probability of failure per demand* (*pfd*). Specifically, the failure process is a Bernoulli process: failures on successive demands are independent events with the same probability (the *pfd*). Bernoulli processes are in common use [1,12]. They give a useful model for many systems where the main concern is design faults, and/or for limited periods of operation. We expect similar results to hold for systems with failure processes in continuous time, and that the approach can be extended to other forms of reliability functions.

To this general picture, the present work adds the aforementioned new scenarios of CII evidence, solutions of the related extremization problems for applying CBI, and illustrations of the impact of the various forms of evidence on confidence in claims. We illustrate how the impact of evidence – on confidence in a claim – can vary significantly, depending on the nature and strength of the evidence. In particular, we highlight situations where additional CII evidence improves confidence in *pfd* bounds, and situations where it does not.

In the rest of this paper, Sec. II discusses related work, and Sec. III reviews CBI. Sec. IV then gives example scenarios in which our models may be applied, and of the different formal CII statements that apply. Sec. V presents the formal statistical models, with illustrative examples shown in Sec. VI. Sec. VII gives a sensitivity analysis, by comparing scenarios that differ in prior knowledge and supporting evidence for CII. Sec. VIII gives a final discussion of our results.

II. Related Work

Bayesian methods are well established in reliability and safety [12]. They support combining various forms of evidence and direct reliability predictions. The results that we use here derive mostly from work on the effects of software faults, or other causes of systematic failure. Software reliability assessment has long been argued to require a statistical approach, and Bayesian methods are well-suited for it [13–15].

Demonstrating that a system meets its dependability requirements, on the sole basis of observed good behaviour (few or zero failures over many demands), is in some cases extremely challenging: it would require observing infeasibly many failure-free runs by the system being assessed, or require improbably strong prior evidence [3,4,11]. However, in many situations, suitable application of Bayesian inference does support strong claims – e.g. situations with more modest reliability requirements, or with justifiable estimates of the probability that certain subsystems will not fail (i.e. a probability of "perfection", or of $pfd = 0$), or with architectural information to support white-box assessment – [6,16,17].

The present work uses conservative Bayesian inference (CBI). This approach is suitable for various safety assessment contexts and produces posterior measures of reliability that are "guaranteed-to-be-conservative", but no more conservative than prior evidence and the observed failure behaviour of the system will allow. For instance, having seen the system successfully handle n demands from its operational environment, CBI gives conservative values for 1) the probability that the system fails on the next demand [9], 2) the probability that the system "survives" the next m demands [10], and 3) the posterior confidence in an upper bound on the system's *pfd* [11]. CBI has also been applied when (rare) failures occur among many correctly handled demands [11,18].

The inference program in CBI applications is the same. An assessor (a) chooses a posterior measure of reliability, (b) specifies an appropriate likelihood function to characterise any observed failure/success behaviour, (c) translates prior evidence into mathematical statements (we will call these "prior knowledge" statements, PKs), (d) considers all prior distributions consistent with these PKs, (e) selects, from this set of priors, a prior that gives the most conservative value for the posterior measure of interest (this need not be unique [7]). For brevity, at times we will use wording like "prior evidence implies a certain effect on a posterior measure", where "prior evidence" needs to be read as "the PKs justified by the prior evidence". The inference program just outlined is closely related to *robust Bayesian analysis* – a general framework for investigating the sensitivity of posterior measures to uncertainties in the inputs of Bayesian inference [19,20].

Unsurprisingly, adding more constraints to limit the set of priors makes CBI's predictions less conservative. Indeed, in the limiting case in which evidence is enough to justify a specific prior distribution, CBI reduces to ordinary Bayesian inference. But even under less extreme circumstances, the extent to which stronger prior evidence can temper CBI conservatism has been studied, e.g., when evidence supporting an estimate of the prior probability of *pfd* being 0 [21], or close to it [10,22], can be included in the assessment.

Similarly, for cases where, as in the present paper, claims for a new system rely on evidence about an older system, we previously showed how justifiable "probability of perfection" evidence can result in less conservative claims than if such evidence were unavailable [7]. We also studied the case of an autonomous vehicle assessed under new environmental

conditions, given knowledge of the system's operation under a different environment in the past [6]. The number of failure-free miles that need to be driven in the new environment – to support a given claim with, say, 90% confidence – can be much less than the number needed to make the same claim if evidence from that previous environment is unavailable.

III. Review: CBI Example

We recall an application of CBI [11]. A Bernoulli process represents the failure behaviour of a system on a succession of demands. Let X be the system's unknown *pfd*. The system is observed to successfully handle n demands. The Bernoulli failure process implies that the probability of observing this sequence of successes – the likelihood function – takes the form $L(x) = (1-x)^n$. Let p be a *pfd* upper bound with respect to which an assessor seeks to make a claim. If the assessor has evidence to support a prior distribution of X, then the posterior confidence that the system *pfd* X is better than p (given that the system survived those n demands) is:

$$P(X \leqslant p \mid n) = \frac{P(X \leqslant p,\ n \text{ successes})}{P(n \text{ successes})} = \frac{\mathbb{E}[\, L(X)\mathbf{1}_{X \leqslant p}\,]}{\mathbb{E}[\, L(X)\,]} \tag{1}$$

where $\mathbf{1}_{\mathtt{S}}$ is an indicator function – it equals 1 when predicate S is true, and 0 otherwise.

But available evidence is typically insufficient to fully justify a given prior for X. Evidence may, instead, support relatively weaker prior claims, such as those proposed below.

Prior Knowledge 1. *certainty that the system pfd X is no better than some $p_l \geqslant 0$. That is, $P(X \geqslant p_l) = 1$.*

Prior Knowledge 2. *$\theta \times 100\%$ confidence that the system pfd X meets, or surpasses, a pfd ε. That is, $P(X \leqslant \varepsilon) = \theta$.*

Here ε, the "engineering goal", is meant as a *pfd* that has high probability θ of being achieved by system developers required to build a system with *pfd* better than p (where $p \geqslant \varepsilon$).

If one has evidence to support prior knowledge 1 and 2, the following proposition (proved in [11]) shows that such knowledge allows one to conservatively gain confidence in a required *pfd* bound p upon observing failure-free operation.

Proposition 1. *Let $\mathcal{D}$ be the set of all probability distributions for the pfd X of a system (i.e. all distributions over $[0,1]$). Consider the optimization problem*

$$\inf_{\mathcal{D}} P(X_B \leqslant p \mid n)$$

(where $\varepsilon \leqslant p$), subject to the constraint that there is evidence the system satisfies prior knowledge 1 and 2.

The prior distribution in Fig. 1 solves this problem because, upon using this prior, $P(X < p \mid n) = \inf_{\mathcal{D}} P(X \leqslant p \mid n)$.

IV. Evidence and Arguments For Improvement Claims

The first, critical step for this form of argument is to examine which evidence supports an "improvement" argument, and translate it into a formal, mathematical CII claim.

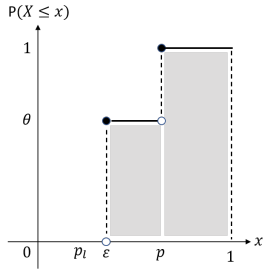


Fig. 1: A conservative prior cumulative distribution function.

This subsidiary claim will support the linking between the reasoning about A and about B, so that evidence collected about A supports confidence regarding B.

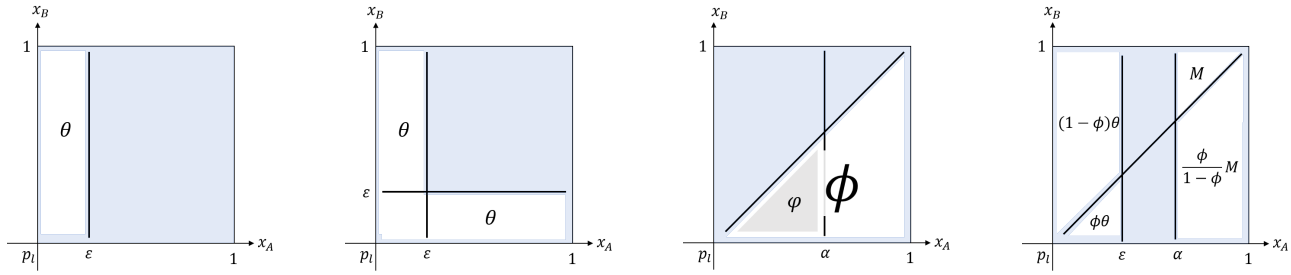
Different evidence may justify different forms of CII claims, thus different final confidence on B. Lack of rigour at this stage could invalidate any conclusions, despite the rest of the reasoning being a provably correct series of deductive steps.

We give some examples of what basis a CII claim may have in evidence; what mathematical form it could then take; and what factors may prevent absolute confidence in the claim, thus requiring that it be considered true only with a certain probability, which we will call ϕ, with $0 < \phi < 1$.

1) A and B are two systems to be used in the same operational environment. B is a newly developed, plug-in replacement for the older system A. B is built to the same specification as A, but by newer methods known to yield better reliability. For instance, its software has been developed with methods known to be less error-prone, and verified through better methods, by better staff. One then expects B to be more reliable than A, as produced by a better technology, thus less likely to have design faults. However, such beliefs concern the generality of systems produced by the two different processes, not A and B specifically: it is possible, though improbable, that B is *worse* than A, as B turns out to be an unusually poor result of a high-quality process; and/or A an unusually good result of the relatively worse process that produced it. These unlikely scenarios determine an amount of doubt $(1-\phi)$.

2) system B is an improvement on system A, identical except that some known defects have been removed (e.g., some failure-prone hardware parts were made more reliable, known design faults were fixed). This would mean that whatever the true reliability of A, B would probably be more reliable in the same environment some of the failures that occur in A due to those defects will not occur in B. This would not be 100% guaranteed of course: bug fixes sometimes introduce new bugs, and occasionally reduce reliability. So, the CII claim that can be made is that whatever the true *pfd* of A, B's *pfd* is better or not worse, with confidence ϕ limited by how often in similar circumstances (system complexity, change approval processes) fixes have actually reduced reliability.

3) system B is obtained by adding to system A some safety protection: e.g., A is a safety system and B adds another independent safety monitor with authority to effect the safety action ("1-out-of-n" scheme). This way of building B ensures



(a) PK1 for A and B; PK2 for A (b) PK1 and PK2 for A and B (c) PK1 and PK3 (d) PK1 and PK4

Fig. 2: How the different forms of prior knowledge that we define constrain joint prior distributions of $\langle X_A, X_B \rangle$, by associating probabilities to regions of the Cartesian plane of (x_A, x_B) values. The forms of prior knowledge are: **(a)** a marginal probability θ that version A meets the engineering goal, $[X_A \leqslant \varepsilon]$; **(b)** identical marginal probabilities θ for both versions; **(c)** probabilities that B is no worse than A, i.e. $P(X_B \leqslant X_A, X_A \leqslant \alpha) = \varphi$ and $P(X_B \leqslant X_A) = \phi$ (where $0 < \varphi \leqslant \phi$ and $\varepsilon \leqslant \alpha$); **(d)** within the events $[p_l \leqslant X_A \leqslant \varepsilon]$ and $[\alpha \leqslant X_A \leqslant 1]$, the ratio between the probabilities for regions above and below the diagonal is $\frac{\phi}{1-\phi}$, for any M, ϕ such that $0 < M < 1-\phi$.

that the set of demands on which B fails unsafely, U_B, is a subset of those where A fails unsafely, $U_B \subseteq U_A$. Hence,

$$pfd_B = \sum_{D \in U_B} P(D) \leqslant \sum_{D \in U_A} P(D) = pfd_A \qquad (2)$$

despite our not knowing either the sets U_B, U_A or the probabilities associated to them. Thus, the evidence supports a claim that *no matter what the true value of* pfd_A, $pfd_B \leqslant pfd_A$. This claim can usually be considered true with very high confidence; however, again, there are conceivable, unlikely scenarios in which adding safety elements increases *pfd*, and the historical frequency of such events will determine the doubt $(1-\phi)$.

4) A is a testing environment, for a demand-operated control system that is to be deployed in an operational environment B. Environment A has been made "stressful" through two precautions: (a) making the statistical distribution of demands (sequences of inputs to the system), conditional on each *type* of demand, the same as in B; but (b) giving higher probabilities to types of demands that are known to cause failures with higher probabilities, due to known limits of the system hardware. Let us call q_i the *pfd* conditional on a demand belonging to type i, for each one of n types of demands; and t_i^A, t_i^B the probabilities of demands of that type in environments A and B respectively. Precaution (a) ensures that no q_i value changes between A and B; and (b) ensures that demand types with higher q_i have higher probabilities in A than in B: for those i, $t_i^A \geqslant t_i^B$. The *pfd* values in the two environments are then as in the equalities below [23]:

$$pfd_B = \sum_{1..n} t_i^B q_i \leqslant \sum_{1..n} t_i^A q_i = pfd_A \qquad (3)$$

where the inequality is due to precaution (b) above. Thus, one can claim, as in case 3, that no matter what the true value of pfd_A is, $pfd_B \leqslant pfd_A$. As before, there is some reason for doubt $(1-\phi)$, e.g. defects in test generation software might violate the invariance of the q_i terms; or the identification of "stressful" demand types may prove wrong.

5) the same scenario as in 4), but the testing environment is made stressful by exaggerating the frequency of demand types that are understood to be more likely to be affected by software design faults: experience with previous systems indicates that the *pfd* tends to be higher for those classes of demands. So, we could reuse equation (3) but replacing the q_i, seen now as random variables, with their expected values. Therefore, pfd_A and pfd_B must also be replaced by their expected values, and the inequality about which the CII claim is made becomes $\mathbb{E}[pfd_B] \leqslant \mathbb{E}[pfd_A]$.

6) A and B are operational environments for a COTS system (e.g. an industrial PLC) which was developed for non-critical applications, hence has little formal evidence to prove its dependability, but has been operated extensively in A, proving very reliable. Environment B presents fewer of the input sequences that are generally known to be "stressful" for this category of products. It is also known that the system has been used in other environments and never been reported to be especially unreliable. Hence, there is confidence that the system will be at least as reliable in B as it was in A; however with some small but non-negligible probability that it will prove less reliable, possibly seriously so. The CII claim that is supported is thus the same as in example 1.

V. Conservative Confidence Bounds On PFD

In this section, we model dependability arguments that incorporate GALE/PIU evidence. In particular, we extend CBI methods to derive conservative confidence bounds on a system's *pfd*, for on-demand systems.

For brevity, we talk about "versions" A and B, irrespective of whether our scenario is 1) system B is a newer version of system A and both operate in identical environments, with GALE evidence gathered from both A and B, or 2) the same system is required to operate in a new environment B different from a previous environment A (so PIU evidence is gathered from the system in environment A).

Failures of A and B occur according to independent Bernoulli processes. Let X_A, X_B be the unknown *pfd*s for versions A and B, with an unknown joint prior distribution of

$\langle X_A,\ X_B\rangle$. Joint prior distributions of $\langle X_A, X_B\rangle$ are depicted in Fig.s 2-5. On the Cartesian plane of (x_A, x_B) values, each figure depicts the partition of the distribution's domain induced by constraints on the distribution ("prior knowledge"), and probability masses associated with subsets and limit points.

After observing n_A, n_B failure-free runs of A and B, one may compute conservative posterior confidence in a claim $[X_B \leqslant p]$ for some required bound p. The Bernoulli processes imply a likelihood function $L(x, y) = (1-x)^{n_A}(1-y)^{n_B}$. We seek conservative values of

$$P(X_B \leqslant p \mid n_A, n_B) = \frac{\mathbb{E}[\, L(X_A, X_B)\mathbf{1}_{X_B \leqslant p}\,]}{\mathbb{E}[\, L(X_A, X_B)\,]} \qquad (4)$$

subject to the prior knowledge an assessor possesses. For all of the scenarios we will consider, prior knowledge 1 applies – i.e. certainty that X_A and X_B cannot be better than some p_l. Prior knowledge 2 (i.e. having $\theta \times 100\%$ confidence a version is no worse than target *pfd* ε) may apply to one or both versions. In addition, evidence may also support one of the following:

Prior Knowledge 3. *confidence in version A's pfd being α or better, and in the B version being an improvement:*

$$P(X_B \leqslant X_A, X_A \leqslant \alpha) = \varphi \qquad (5)$$

where $\varepsilon \leqslant \alpha \leqslant 1$ and $0 < \varphi < 1$. In particular, ϕ is defined as the value of φ when $\alpha = 1$ (see Fig. 2c).

Prior Knowledge 4. *confidence in version A's pfd falling within some range of values, and version B being an improvement: for some sub-interval I of $[0, 1]$, with ϕ as just defined,*

$$P(X_B \leqslant X_A, X_A \in I) = \frac{\phi}{1-\phi} P(X_A < X_B, X_A \in I) \qquad (6)$$

In particular, we consider the case when (6) holds for the two intervals $[p_l \leqslant X_A \leqslant \varepsilon]$, $[\alpha \leqslant X_A \leqslant 1]$ and, thus (as probabilities must add up to 1), also holds for $[\varepsilon < X_A < \alpha]$ (see Fig. 2d).

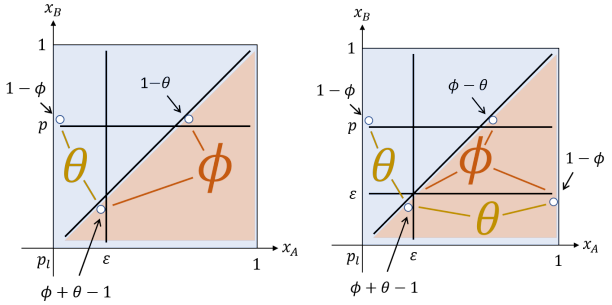
We will refer to either of these two forms of prior knowledge as "*confidence in improvement*" (CII) and, more generally, we will use "PK" to refer to prior knowledge.

PK 1, 2, 3 and 4 impose increasingly stringent constraints on the joint prior distribution of $\langle X_A,\ X_B\rangle$ (see Fig. 2).

Equations (5) and (6) are alternative, but related, CII formalisations that may be supported by: 1) GALE evidence, or 2) arguments justifying why PIU evidence (i.e. n_A from version A) can inspire confidence in version B. They differ in that PK 3 enforces $P(X_B \leqslant X_A, X_A \leqslant \alpha) = \varphi$, while PK 4 is equivalent to enforcing $P(X_B \leqslant X_A | X_A \in I) = \phi$ for 3 choices of the interval I: any one of the 3 "X_A"-intervals stated in PK 4. By definition, $\varphi \leqslant \phi$, with $\varphi = \phi$ when $\alpha = 1$. So, when $\alpha = 1$ and I is the single interval $p_l \leqslant x_A \leqslant 1$, PK 3 and PK 4 agree and imply $P(X_B \leqslant X_A) = \phi$.

If (6) (PK 4) holds for *any* sub-interval I, this is a stronger constraint than either PK 3 or PK 4: it implies B being an improvement (over A) is statistically independent of how reliable A is – i.e. $P(X_B \leqslant X_A \mid X_A \in I) = \phi = P(X_B \leqslant X_A)$ for any I (examples 3, 2 in section IV). Observing failure-free runs of A will not alter confidence in B being an improvement, i.e., $P(X_B \leqslant X_A \mid n_A) = \phi$. See [24] for proofs.

The problem of determining conservative values of (4) now becomes a collection of constrained optimization problems, each subject to varying forms of PK. In each case, the optimization is over a constrained subset of $\mathcal{D}$, where $\mathcal{D}$ is the set of all prior probability distributions of $\langle X_A, X_B\rangle$. And each prior that solves the optimization is referred to as "the most conservative" prior, in the sense that for these priors, $P(X_B < p \mid n_A, n_B) = \inf_{\mathcal{D}} P(X_B \leqslant p \mid n_A, n_B)$. We now state these optimization problems as three propositions.



(a) $\phi > 1-\theta$ (b) $\phi > \theta > 1-\theta$

Fig. 3: Prior distributions that solve proposition 2, when evidence supports **(a)** PK 2 for only version A, or **(b)** PK 2 for both versions (for the parameter ranges given in the respective subcaptions).

Proposition 2 generalises a result in [11]. Its proof mirrors those for propositions 3 and 4 below, outlined in [24].

Proposition 2. *Consider the optimization problem*

$$\inf_{\mathcal{D}} P(X_B \leqslant p \mid n_A, n_B)$$

(where $\varepsilon \leqslant p$), subject to the constraints that there is evidence:

1) A and B satisfy PK 1, 3, with $\alpha = 1$ and $\phi > 1-\theta$;
2) either A alone, or A and B, satisfy PK 2.

The prior distributions in Fig.s 3a and 3b solve this problem for certain parameter ranges of the constraints.

Here, the assessor believes $P(X_B \leqslant X_A) = \phi$; this is what (5) means when $\alpha = 1$. Fig.s 3a and 3b represent cases when this confidence ϕ is high: i.e. $\phi > \theta$ and $\phi > 1-\theta$. Due to this strong CII, these priors (out of all that solve proposition 2) give greatest posterior confidence in $[X_B \leqslant p]$.

The following two propositions are novel (proofs in [24]).

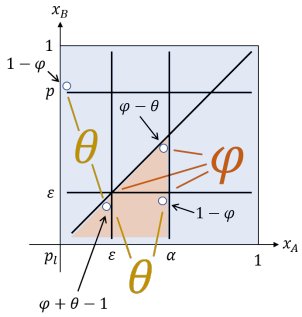
Proposition 3. *Consider the optimization problem*

$$\inf_{\mathcal{D}} P(X_B \leqslant p \mid n_A, n_B)$$

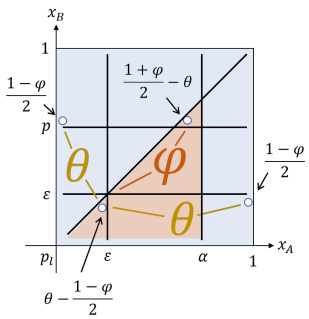
(where $\varepsilon \leqslant p$), subject to the constraints that there is evidence versions A and B satisfy PK 1, 2, 3.

Fig. 4 shows prior distributions that solve this problem for certain parameter ranges of the constraints.

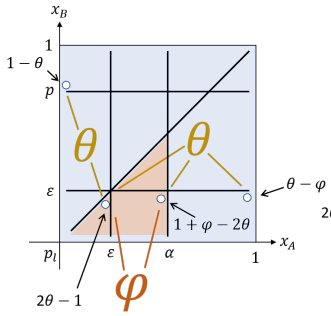
Here, the confidence φ is about $[X_B \leqslant X_A \leqslant \alpha]$. Analogously to the priors in Fig. 3, Fig.s 4a and 4b solve



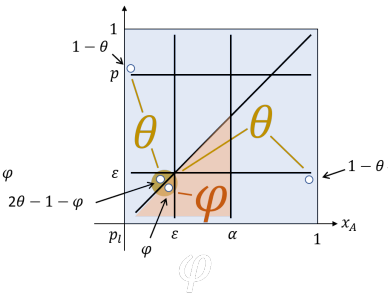
(a)
$\varphi > \theta > 1-\theta,\ \varepsilon < \alpha \leqslant p$



(b)
$\varphi > \theta > 1-\theta,\ p < \alpha < 1$

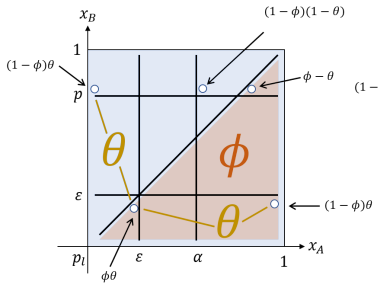


(c)
$\theta > \{1-\theta,\ \varphi\}$
$\theta - \varphi \geqslant 1-\theta,\ \varepsilon < \alpha < 1$

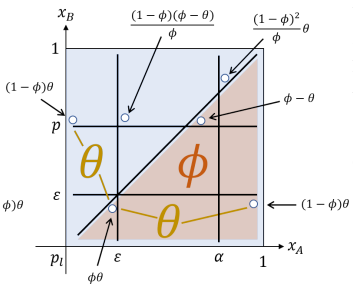


(d)
$\theta > \{1-\theta, \varphi\},\ \theta - \varphi \leqslant 1-\theta$
$\varepsilon < \alpha < 1$

Fig. 4: Prior distributions that solve proposition 3, depending on the value of α and the strength of evidence supporting the proposition's constraints (i.e. the relative sizes of parameters in Fig.s 2b and 2c).



(a) $\phi > \theta,\ \varepsilon < \alpha \leqslant p$



(b) $\phi > \theta,\ p < \alpha < 1$

Fig. 5: Prior distributions that solve proposition 4, depending on α and the strength of evidence supporting the proposition's constraints (i.e. the relative sizes of parameters in Fig.s 2b and 2d).

proposition 3 when $\varphi > \theta > 1-\theta$ and give the greatest posterior confidence among the priors in Fig. 4. Note that 4b, 4c and 4d all possess B-version marginal distributions identical to the single-system CBI prior of proposition 1 (i.e. Fig. 1).

Proposition 4. *Consider the optimization problem*

$$\inf_{\mathcal{D}} P(X_B \leqslant p \mid n_A, n_B)$$

(where $\varepsilon \leqslant p$), subject to the constraints that there is evidence versions A and B satisfy PK 1, 2, 4.

Fig. 5 shows prior distributions that solve this problem for certain parameter ranges of the constraints.

Again, analogously to the priors in Fig.s 4a and 4b, the priors in Fig.s 5a and 5b solve proposition 4 when $\phi > \theta$, i.e. evidence strongly supports $[X_B \leqslant X_A]$.

VI. Numerical Illustrations

The conservative claims derived in Sec. V can be applied in scenarios with evidence to support the related PKs. These claims are summarised as formulae in the last column of table I. Each formula is $P(X_B < p \mid n_A, n_B)$, computed from the conservative prior figure indicated in the table row. In [24], we prove this is the greatest lower bound on $P(X_B \leqslant p \mid n_A, n_B)$.

Consider the following two illustrations.

Example 1. *Consider a nuclear reactor protection software (which is simple enough to possibly be perfect [16], thus $p_l = 0$)[1] whose old version A has been exposed to $n_A = 100$ demands without failures in previous operation of the nuclear reactor. Now a new version B is believed to be more reliable (with confidence ϕ) due to, e.g., employing the same basic design methods but with more advanced formal verification techniques. For each version, the assessor has a high $\theta \times 100\%$ confidence that the engineering goal $\varepsilon = 10^{-6}$ has been achieved. So, upon observing no failures in n_B demands during operational testing of B, the assessor's conservative confidence $c \times 100\%$ in claim $[X_B \leqslant 10^{-4}]$ is shown in Fig. 6.*

Fig. 6 shows how stronger CII supporting evidence can result in greater confidence c in a *pfd* bound, and such benefit is more obvious when the prior confidence θ in the engineering goal is relatively weaker. That is, when $\theta = 0.7$, increasing ϕ from 0.8 to 0.99 results in a greater improvement in confidence than when $\theta = 0.9$ – the gap between the dotted blue curve and the dashed green curve is bigger than the gap between the dash-doted red curve and the solid orange curve.

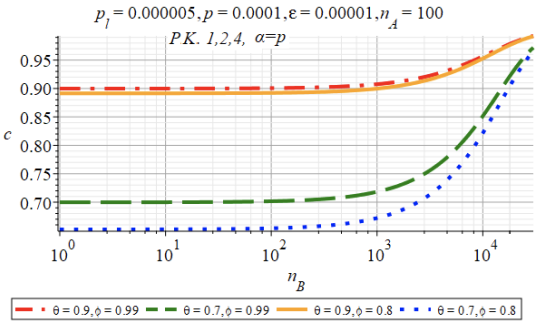


Fig. 6: Example 1. Posterior confidence c in $[X_B \leqslant 10^{-4}]$ for various θ and ϕ, as a function of failure-free runs of version B.

The following example uses parameters obtained from [6].

Example 2. *An autonomous vehicle (AV) has an unknown probability of fatality-event per mile (pfm) – an analogue of pfd. The engineering goal, $\varepsilon = 10^{-10}$, is 2 orders of magnitude*

[1] It normally forms one channel of a 1-out-of-n protection system [16,22]

TABLE I: Conservative estimates of $P(\, X_B \leqslant p \mid n_A, n_B \,)$ supported by prior knowledge 1, 2, 3 or 4. Except where explicitly stated otherwise, prior knowledge 1 and 2 apply to both versions.

prior knowledge	conservative priors	$\inf_{\mathcal{D}} P(X_B \leqslant p \mid n_A, n_B))$
PK 1, 2 (only version A) & 3 (see Fig.s 2a and 2c), $\varphi = \phi > 1-\theta$	Fig. 3a	$\frac{(\phi+\theta-1)L(\varepsilon,\varepsilon)}{(\phi+\theta-1)L(\varepsilon,\varepsilon)+(1-\phi)L(p_l,p)+(1-\theta)L(p,p)}$
PK 1, 2 & 3 (see Fig.s 2b and 2c), $\varphi = \phi > \theta > 1-\theta$	Fig. 3b	$\frac{(\phi+\theta-1)L(\varepsilon,\varepsilon)}{(\phi+\theta-1)L(\varepsilon,\varepsilon)+(1-\phi)L(p_l,p)+(\phi-\theta)L(p,p)}$
PK 1, 2 & 3 (see Fig.s 2b and 2c), $\varepsilon \leqslant \alpha \leqslant 1$	Fig. 4a	$\frac{(\varphi+\theta-1)L(\varepsilon,\varepsilon)+(1-\varphi)L(\alpha,\varepsilon)+(\varphi-\theta)L(\alpha,\alpha)}{(\varphi+\theta-1)L(\varepsilon,\varepsilon)+(1-\varphi)L(\alpha,\varepsilon)+(\varphi-\theta)L(\alpha,\alpha)+(1-\varphi)L(p_l,p)}$
	Fig. 4b	$\frac{(\varphi+2\theta-1)L(\varepsilon,\varepsilon)}{(\varphi+2\theta-1)L(\varepsilon,\varepsilon)+(\varphi+1-2\theta)L(p,p)+(1-\varphi)L(p_l,p)}$
	Fig. 4c	$\frac{(2\theta-1)L(\varepsilon,\varepsilon)+(1+\varphi-2\theta)L(\alpha,\varepsilon)}{(2\theta-1)L(\varepsilon,\varepsilon)+(1+\varphi-2\theta)L(\alpha,\varepsilon)+(1-\theta)L(p_l,p)}$
	Fig. 4d	$\frac{(2\theta-1)L(\varepsilon,\varepsilon)}{(2\theta-1)L(\varepsilon,\varepsilon)+(1-\theta)L(p_l,p)}$
PK 1, 2 & 4 (see Fig.s 2b and 2d), $\varepsilon \leqslant \alpha \leqslant 1$	Fig. 5a	$\frac{\phi\theta L(\varepsilon,\varepsilon)}{\phi\theta L(\varepsilon,\varepsilon)+\theta(1-\phi)L(p_l,p)+(1-\phi)(1-\theta)L(\alpha,p)+(\phi-\theta)L(p,p)}$
	Fig. 5b	$\frac{\phi^2\theta L(\varepsilon,\varepsilon)}{\phi^2\theta L(\varepsilon,\varepsilon)+\phi\theta(1-\phi)L(p_l,p)+(1-\phi)(\phi-\theta)L(\varepsilon,p)+\phi(\phi-\theta)L(p,p)+\theta(1-\phi)^2L(\alpha,\alpha)}$

safer than the pfm for human drivers[2]*. The risk of catastrophic hardware failures implies that* $p_l = 10^{-15}$*. The AV company, upon testing the AV in City-A for* n_A *fatality-free miles, wants to deploy in City-B. The company is confident* ($\theta = 0.9$) *that the AV performs no worse than* ε *in each city. And, they have some confidence* ϕ *that the road/weather conditions of B are similar or more favourable, so this environmental change should not harm safety. What conservative number of new fatality-free miles* n_B *need to be driven in B to claim – with* 95% *confidence – that the AV is as safe as the average pfm for human drivers,* 10^{-8} *(i.e. the claim* $[X_B \leqslant 10^{-8}]$*)? The answer is presented in Fig. 7 as a function of* n_A*.*

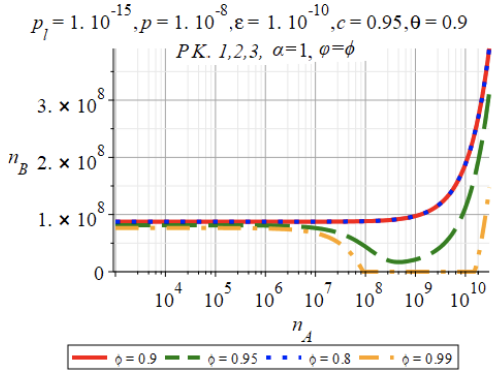


Fig. 7: Example 2. Number of failure-free runs n_B, given n_A failure-free runs for A, required to achieve 95% confidence in $[X_B \leqslant 10^{-8}]$.

In Fig. 7, stronger CII supporting evidence (i.e. increasing ϕ) may reduce the required n_B, or may not – depending on whether $\phi \geqslant \theta > 1-\theta$ (e.g. $\phi = 0.95$ in prior Fig. 3b) or $\theta \geqslant \phi > 1-\theta$ (e.g. $\phi = 0.8$ in the $\alpha \to 1$ limit of either prior Fig. 4c or 4d). In fact, notice that the relevant formulae for both Fig. 4c and 4d (see in table I) give $\frac{(2\theta-1)L(\varepsilon,\varepsilon)}{(2\theta-1)L(\varepsilon,\varepsilon)+(1-\theta)L(p_l,p)}$ in the limit. This clearly does not depend on ϕ, which is why the solid curve and dotted curve in 7 are identical.

How much increasing CII reduces n_B depends on n_A. Starting from $n_A = 0$, the more n_A are observed, the fewer the n_B needed to support posterior confidence in the claim – n_B may even reach zero. But eventually, as n_A increases, n_B increases without bound. This is because the stated PK does not exclude the possibility that B is very unreliable if A is very reliable. In fact, with increasing n_A the posterior probability of the (x_A, x_B) point (p_l, p) – an undesirable point from a safety viewpoint – grows arbitrarily close to 1, requiring an arbitrarily large n_B to relocate probability to more desirable points (e.g. to $(\varepsilon, \varepsilon)$ below the $[X_B = p]$ horizontal line) and improve posterior confidence in the 10^{-8}, X_B bound. All of the worst-case priors in Sec. V allow this effect (in line with previous observations [6]).

VII. Sensitivity of Conservative Claims to Alternative Forms/Strengths of Evidence

This section highlights the change in conservative confidence claims, in response to changes in the strength of supporting dependability evidence and the PKs. A useful reference scenario is the assessment of a single system (i.e. proposition 1), where 95% posterior confidence (in a claimed *pfd* upper bound p) is supported by a PK 2 confidence of $\theta \times 100\%$. In particular, if $p = 10^{-8}$ and $\theta = 0.9$, the most conservative prior Fig. 1 implies that $n = 7.55e7$ failure-free runs of this system are needed to support the claim [11].

[2]The exact statistic in the U.S. (2013) is 1.09e–8, as used by [5], while for simplicity we round this to 10^{-8} in our example.

Analysis 1. *An analysis of the extent to which CII supporting evidence can temper conservatism – in particular, reduce the number of failure-free n_B runs needed to support 95% posterior confidence in the claim $[X_B \leqslant p]$ – compared with if alternative evidence is used. We compare n_B values related to the following alternative CII formalisations and PK:*

1) *PK 1, 2, 3, $\varphi = \theta\phi$, $\alpha = p$, (prior Fig. 4d);*
2) *PK 1, 2, 3, $\varphi = \phi$, $\alpha = 1$, (prior Fig. 3b);*
3) *PK 1, 2, 4, ϕ, $\alpha = p$, (prior Fig. 5b);*
4) *PK 1, 2, 3, $\varphi = 0.95$, $\alpha = p$, (prior Fig.s 4a, 4b);*
5) *PK 1, 2 (for version A), 3, $\varphi = \phi$, $\alpha = 1$, (prior Fig. 3a).*

Above, $\phi = 0.95$, $\theta = 0.9$ and $p = 10^{-8}$. And, PK 2 applies to both versions, except where explicitly stated otherwise.

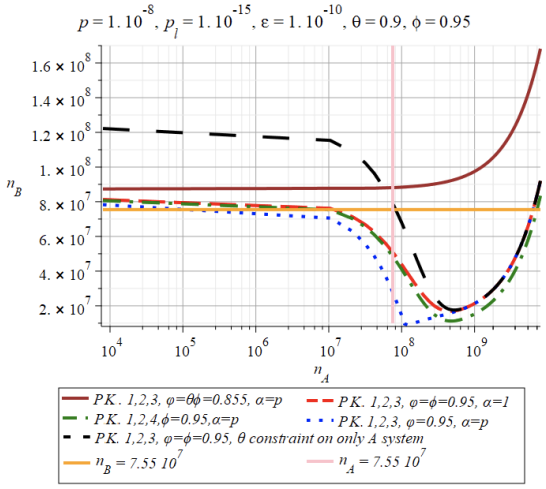


Fig. 8: The forms of CII and PK, the strength of supporting evidence (i.e. the values of θ, ϕ, φ), and n_A evidence, all matter (to varying degrees) in reducing the number of failure-free runs n_B needed to support 95% posterior confidence in the claim $[X_B \leqslant 10^{-8}]$.

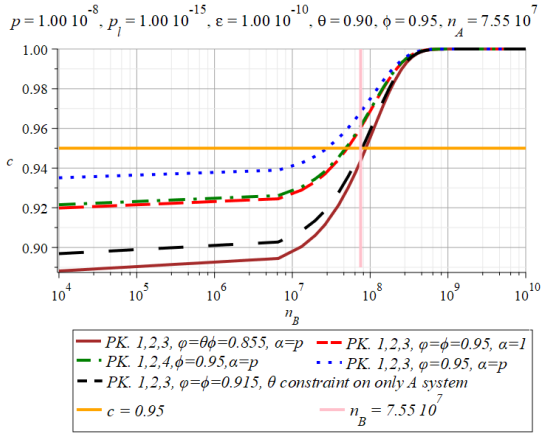


Fig. 9: Stronger forms of CII give greater confidence c in claims.

Asymptotically, for large n_A, all the priors require very large n_B to support $[X_B \leqslant 10^{-8}]$ (Fig. 8). Because, without evidence to the contrary, it is conservative to believe that version B is extremely unreliable *when* version A is extremely reliable. None of the PK represent such contrary evidence.

Also, for small n_A, the curves all lie above $n_B = 7.55e7$ (Fig. 8). Thus, the n_B each prior needs to support the claim is more runs than is needed in the single system reference scenario. This is despite being very confident that B is an improvement (e.g. $\phi = 0.95$). Because, after observing *only one successful run* from A, all these priors have posterior X_B distributions that give less confidence than the single system prior of Fig. 1. Without evidence to the contrary, it is conservative to believe version B is extremely reliable *when* version A is extremely unreliable. Again, none of the PK represent such contrary evidence.

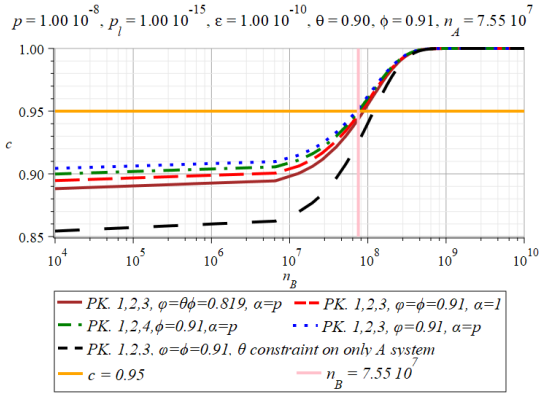


Fig. 10: Weaker CII supporting evidence ($\phi = 0.91$) gives weaker confidence c in claims (compared to Fig. 9), but to varying extents depending on the form of CII.

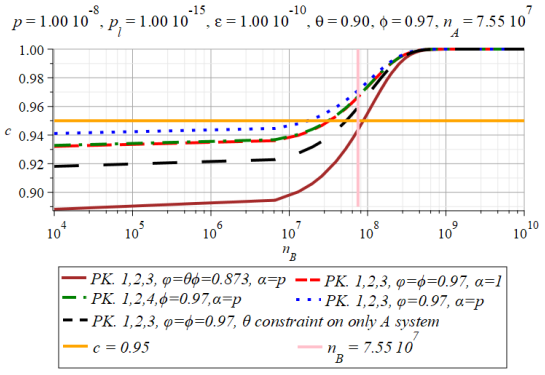


Fig. 11: Stronger CII supporting evidence ($\phi = 0.97$) gives stronger confidence c in claims (compared to Fig. 9), but to varying extents depending on the form of CII.

Notice how there is a changing ordering of the priors, in terms of how many n_B runs they need to support the claim. Initially, the prior with the strongest CII evidence in PK 3 (i.e. $\varphi = 0.95$) requires the fewest n_B (dotted curve), while the prior with PK 2 satisfied by only the A version requires the most n_B (dash-space curve). Eventually however, the prior with PK 4 evidence requires the fewest n_B (dash-dot curve). While the prior with the weakest PK 3 evidence (i.e.

$\varphi = \phi\theta$) requires the most n_B (solid curve) – here, statistical independence (if true) gives a value for φ in terms of θ and ϕ. Clearly, such statistical independence eventually leads to very conservative n_B requirements.

Fig. 9 is an alternative view for analysis 1. If $n_A = 7.55e7$ – i.e. equal to the failure-free runs needed in the single system scenario, to support 95% confidence in a 10^{-8} *pfd* bound – then some forms of CII give more confidence than others. In this sense the CII forms are ordered, from smallest to greatest confidence, as PK 3 (with $\alpha = 1$), 4 and 3 (with $\alpha < 1$).

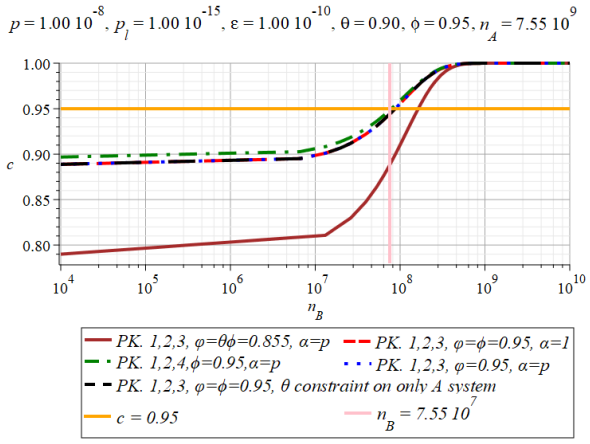


Fig. 12: Increasing n_A evidence leads to greater confidence c in claims (compared to Fig. 9), but not in all cases.

Fig. 10 shows, by comparison with Fig. 9, how weaker CII-supporting evidence reduces confidence in claims. Reducing ϕ from 0.95 to 0.91 noticeably reduces confidence in most claims. However, the inference that uses the independence assumption $\varphi = \phi\theta$ shows no change over Fig. 9. This is because the independence guarantees that $\theta > \varphi$ (since $\phi < 1$), and the probabilities are such that $\theta - \varphi < 1 - \theta$. So that the relevant worst-case prior is Fig. 4d (for $\alpha = p$), which gives posterior confidence that *does not depend on* φ (see relevant formula for Fig. 4d in table I). Analogously, Fig. 11 (with $\phi = 0.97$) demonstrates how stronger CII evidence leads to greater confidence in claims, compared to Fig. 9. Confidence in almost all the claims has increased noticeably; the extent of this increase depends on the form of CII.

We remarked on how all the priors in this paper require very large n_B when n_A is very large (say $n_A > 10^{10}$). Fig. 12 depicts what happens when $n_A = 7.55e9$, i.e. 2 orders of magnitude more than $n_A = 7.55e7$ in Fig. 9. All of the priors now give smaller posterior confidence, their respective curves lying everywhere lower than before. But this drop in confidence happens at different rates for the different priors. Notice, the PK 4 prior now gives greater posterior confidence than a PK 3 prior (with $\varphi = 0.95$), where it did not before.

Contrastingly, for $10^7 < n_A < 10^9$, most of the priors show less conservatism (i.e. smaller required n_B) as n_A increases (Fig. 8). And, for some priors, there exist unique n_A values such that, with this failure-free "A" evidence, the posterior confidence from the priors cannot be higher. Analysis 2 compares how much increase in confidence these n_A bring.

Analysis 2. *An analysis of the increase in confidence that n_A can bring. Consider the priors with the following PK. For each prior, the respective confidence formula in table I determines n_A^* – the unique n_A for the prior such that no other n_A supports higher posterior confidence in $[X_B \leqslant p]$:*

1) *PK 1, 2 (only version A), 3, $\varphi = \phi$, $\alpha = 1$, $n_A^* = 5.29e8$ (prior Fig. 3a);*
2) *PK 1, 2, 3, $\varphi = \phi$, $\alpha = 1$, $n_A^* = 4.60e8$ (prior Fig. 3b);*
3) *PK 1, 2, 4, ϕ, $\alpha = p$, $n_A^* = 4.75e8$ (prior Fig. 5b).*

Above, $\phi = 0.95$, $\theta = 0.9$ and $p = 10^{-8}$. And, PK 2 applies to both versions, except where explicitly stated otherwise.

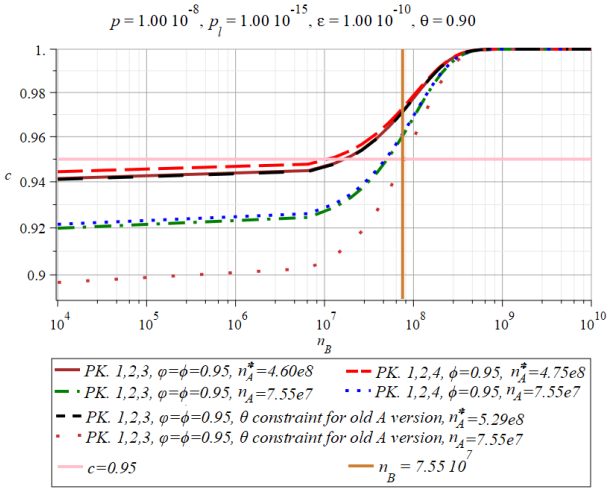


Fig. 13: The increase in confidence c, due to n_A^* evidence, depends on whether PK 2 is satisfied by both versions, and on the CII form.

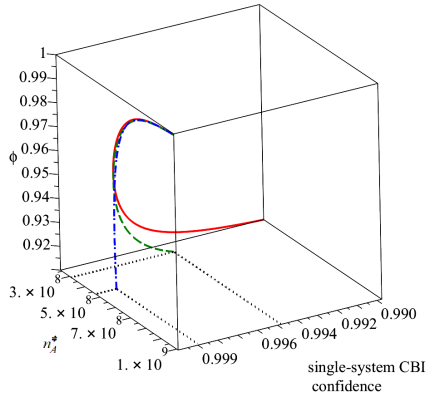


Fig. 14: Confidence in the "single system" claim $[X \leqslant 10^{-8}]$ upon observing n failure-free runs of the system, if $n = n_A^*$. Here, n_A^* is a function of ϕ, for prior Fig.s 3a (dash-dot curve), 3b (solid curve), and 5b (dash curve) of analysis 2.

Fig. 13 shows the greatest amount of posterior confidence each of these priors can give – the unique n_A^* for each prior

ensures these upper limits are reached *for all n_B failure-free evidence from the* B *version*. If claims were supported by n_A^* rather than, say, the reference $n_A = 7.55e7$, then the largest resulting increase in confidence is experienced by prior Fig. 3a. This prior is supported by "engineering goal" confidence PK 2 for only the A version. For the other two priors, PK 2 is satisfied by both versions. This suggests that the additional $\theta \times 100\%$ confidence (in B satisfying the engineering goal) also brings a noticeable increase in posterior confidence in claims, even when $n_A \neq n_A^*$. Also note that, unlike the change in ordering of Fig. 12 due to large n_A, the n_A^*s do not change the ordering of the priors in Fig. 9 – so, among these priors, prior Fig. 5b still gives the greatest confidence.

For each prior in analysis 2, n_A^* is a function of the strength of CII supporting evidence (i.e. n_A^* is a function of ϕ). Very large n_A^* accompanies very strong CII evidence (Fig. 14). In fact, if these n_A^* number of failure-free runs were observed in the single system reference scenario, the posterior confidence in a 10^{-8} bound on *pfd* would be upwards of 99% (Fig. 14)!

VIII. Discussion & Conclusion

A. Why CBI; risk of spurious optimism

Assessing system dependability may require integrating diverse forms of evidence. For an assessor presented with such evidence, Bayesian methods are a principled statistical toolkit for dealing with uncertainty. However, they bring the challenge of specifying a suitable prior distribution – one that fully captures, and captures only, the assessor's beliefs about how reliable the system might be, as justified by evidence. A prior may be wrong in that it fails to capture all of an assessor's beliefs. Or because it encodes additional beliefs not actually held by the assessor. A wrong prior could, unbeknown to the assessor, lead to dangerously optimistic assessments, or unduly undermine confidence in a good system.

In this paper we have made this task of specifying priors even harder by making claims for B depend, in part, on evidence about A.

We addressed this problem via *conservative Bayesian inference* (CBI). CBI's advantages include: 1) it encourages assessors to be minimalist, i.e., specify only those beliefs which can be justified by the evidence; 2) it produces provably conservative claims (no other prior distribution satisfying the beliefs will yield more conservative claims); 3) CBI allows its users to see how much unjustified confidence would be added by using an individual prior, among the priors allowed by the PKs, instead of the worst case; 4) by spelling out worst-case distributions compatible with the stated beliefs, CBI highlights apparent inadequacies in the beliefs stated, prompting assessors to review how well the stated beliefs reflect the evidence.

The prior distributions of Fig.s 3, 4 and 5 each give the least confidence in an upper bound, p, on B's *pfd*, depending on various forms (and strengths) of reliability evidence. This evidence includes: observed failure-free operation on sets of n_A and n_B demands; evidence justifying CII probabilities ϕ, φ of B being an improvement over A, or justifying the probability θ of a version's *pfd* being at least as good as the engineering goal ε.

B. Selection of formal PKs and of parameter values

The formal probabilistic characterisations of CIIs that we have introduced in the form of PKs 3 and 4 cover a set of scenarios commonly found in practice. This substantially broadens the set of scenarios we studied previously [6,7].

The forms of CII (and hence PKs) to use in a CBI safety argument should be determined by careful consideration of prior evidence (as we exemplified in Sec. IV). Reasonable values for φ or ϕ would normally be suggested by prudent use of historical evidence about similar systems.

How to translate available prior evidence into formal PKs will not always be obvious. In some cases, evidence may seem to support more than one form of PK. Comparing and contrasting their implications may then be useful, revealing gaps in one's understanding of what the evidence implies. This exploration can be helped by the fact that sometimes parameter values related to one form of CII can be derived from parameter values for another CII form. E.g., if evidence justifies a probability of A satisfying the required *pfd* bound, $P(X_A \leqslant p) = 1-\gamma$, and given PK 3 with $\alpha = p$, one can then assign φ the value $\phi - \gamma$, the smallest (i.e. most conservative) value of φ consistent with ϕ and γ. Here, γ could be the result of previously applying CBI to version A (Sec. III).

We showed (Sec. VII) how some forms of PK result in more confidence in a claim than other forms; and some require fewer failure-free demands than others in order to support a given claim. We emphasize that one must not use this knowledge for simply claiming that PK that yields the most favourable results – an unethical and dangerous practice. The choice of PKs must be based on the prior evidence alone.

In certain cases, obtaining evidence that might bring added confidence may be expensive – for instance, if discovering what one's φ value should be would require analysis of extensive logs about past projects. A decision is then needed about whether the added confidence justifies the significant cost. Our results can inform such considerations.

We observe that parameter values sometimes need not be specified with great precision. E.g., (see Fig. 7): if CII supporting evidence is relatively weak (i.e. $\theta > \phi > 1 - \theta$) – i.e. confidence in B being an improvement is less than confidence that A and B meet their "engineering goal" – then posterior confidence, and the n_A and n_B needed to support a claim, do not depend on ϕ.

We have focused on those CBI priors (parameter values for the PKs) consistent with what we would expect in practice. For instance, we assumed confidence higher than 50% (i.e. $\theta > 1 - \theta$) that the engineering goal is met: system development would not usually be started if confidence were lower.

We have also not considered the case of $p < \varepsilon$, i.e, of a required bound on *pfd* lower than the value one has some confidence of being achieved. Previous studies showed how this may give zero posterior confidence unless additional prior knowledge is stated, about values of *pfd* smaller than p [7,9].

A broader observation is appropriate about the worst-case priors that CBI produces. Recall, for instance, how in Fig. 8 increasing n_A eventually undermines confidence in a claim, requiring increasing n_B to compensate. This is because the worst case priors include a probability mass at point $(p_l,\, p)$; that is, a belief "if A turns out to be extremely good, then B must be inadequate, but just inadequate enough not to achieve the desired *pfd*, p". We trust that readers will rightly object to such a belief, because experience would not typically support it. And yet, such a belief is not refuted by any of the forms of PK 1 to 4 that an assessor could specify to support a safety claim. We now discuss the implications of such apparently unreasonable worst-case priors.

If one followed the steps we recommended, proceeding from the evidence to carefully spell out which PKs it implies, and yet the resulting worst-case prior seems absurd, possible reasons are: 1) the analysis may have been inadequate, and requires additional PK that capture neglected implications of the evidence [9,21,22]; 2) that prior, albeit absurd, may be a limiting case of a class of priors which are themselves plausible, in which case we cannot rightly forbid it just because it leads to unpleasant conclusions; 3) the "absurdity" of the results flags an error we made in deriving the PKs from the evidence, or in our likelihood function, etc.

Sec. VII also highlights how, unsurprisingly, strongly supported CII can increase confidence in posterior claims, or reduce the n_B needed for a stated level of confidence. However, when n_A is small, *all* of the worst-case priors require a larger n_B than required when making the same claim on B using a "single-system" CBI prior (see discussion of Fig. 8).

We also highlight how $\theta \times 100\%$ confidence for both versions brings a noticeable increase in posterior confidence – i.e. the difference between the prior from Fig. 3a (only A version) and the prior from Fig. 3b (both versions) in Fig. 13. This is true when n_A is not the value n_A^* that guarantees the greatest posterior confidence. If $n_A = n_A^*$, θ for both versions brings no benefit; what seems to matter then is the form of CII. Knowing n_A^* tells us how much more confident, at most, a claim for B could be, everything else being equal.

Finally, unlike ordinary Bayesian inference, in CBI the form of prior used needs to depend on the input values. For example, up until about $n_A = 10^8$, the dotted curve in Fig. 8 is generated by the prior in Fig. 4b. Beyond that point, it is generated by the new worst-case prior in Fig. 4a. These priors are very different, and erroneously using only one of these *for all* n_A would give significantly over-optimistic results.

C. Future Work

The forms of formal "prior knowledge" that we have studied are chosen to be realistic, but do not exhaust those that may hold in practice. Case studies, especially applying the kind of scrutiny we have outlined above when CBI produces apparently unreasonable worst-case priors, may reveal other forms of prior evidence that can reduce excess conservatism.

In this direction – addressing over-conservative worst-case priors – a purely mathematical next step is to find a solution that fully exploits the form of claim in example 3 (Sec. IV), requiring (6) to hold for *every* subinterval in $[0, 1]$. Using this form of CII to address example 3 would give more confidence in claims on *pfd*, compared with approximating it by PK 4.

We have focused on scenarios in which no failures occurred. This makes sense for certification in critical systems, regarding systematic failures: usually, if a software failure occurs, the system is fixed, and reassessed from scratch. If this reassessment ignores that the fault in the previous version may undermine assumptions on which the assessment relies, its result may be over-optimistic [25]. Accordingly, it would be useful to extend the present work (similarly to [25]), to assessing version B given failures in version A; or, for less critical systems, given failures in both A and B.

Some aspects of the mathematical apparatus can be easily completed if needed. For instance, we do not explicitly account for the joint probability of both versions satisfying the engineering goal. There may be reasons for believing either positive or negative correlation between the two events. Studying their effects may give more insight into useful forms of PK that are currently missing. Also, in our scenarios, the same $\theta \times 100\%$ confidence applies to both versions. The two could be different in practice (e.g. due to different approaches applied in developing versions A and B, or markedly different operational environments).

Thus far, CBI applications have involved solving constrained mathematical optimizations over sets of prior probability distributions. This captures the uncertainty an assessor has in adequately specifying what beliefs prior evidence justifies. But the assessor could also have uncertainty in specifying the probabilistic failure model for the failure-free observations (i.e., the likelihood function). Extending CBI to assess the effects of such uncertainty would be a fruitful exercise.

D. Summary of contributions of this paper

In this paper, we have reported: 1) how various practical scenarios map into formal "prior knowledge" (PK) statements, to use in conservative Bayesian inference (CBI); 2) convenient closed form solutions for the worst-case priors and for the posterior confidence in claimed *pfd* bounds; 3) sensitivity analyses, identifying parameter ranges for which evidence from operation in a system, or environment, A, reduces the amount n_B of failure-free operation required in system, or environment, B for a required confidence in a bound.

Together with our previous work [6,7], by introducing new examples based on practical scenarios, this paper demonstrates how CBI can be used to formalise arguments that use claims like "proven in use" (PIU), "Globally at least equivalent" (GALE), or "stress tested" – claims derived from operation/test evidence on related, but not identical, environments of use or system versions. These examples are a good guide when translating other forms of prior evidence into formally stated CBI constraints ("PK"s).

ACKNOWLEDGMENT

We thank the anonymous reviewers for their insightful comments and helpful suggestions for improving the paper.

REFERENCES

[1] IEC, *IEC61508, Functional Safety of Electrical/ Electronic/Programmable Electronic Safety Related Systems*, 2010. [Online]. Available: https://webstore.iec.ch/publication/22273

[2] CENELEC, *EN 50129:2018: Railway applications - Communication, signalling and processing systems - Safety related electronic systems for signalling*. European Committee for Electrotechnical Standardization (CENELEC), Nov. 2018.

[3] B. Littlewood and L. Strigini, "Validation of ultra-high dependability for software-based systems," *Comm. of the ACM*, vol. 36, pp. 69–80, 1993.

[4] R. W. Butler and G. B. Finelli, "The infeasibility of quantifying the reliability of life-critical real-time software," *IEEE Transactions on Software Engineering*, vol. 19, no. 1, pp. 3–12, Jan. 1993.

[5] N. Kalra and S. Paddock, "Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability?" *Transp. Research Part A: Policy and Practice*, vol. 94, pp. 182–193, 2016.

[6] X. Zhao, K. Salako, L. Strigini, V. Robu, and D. Flynn, "Assessing safety-critical systems from operational testing: A study on autonomous vehicles," *Information and Software Technology*, vol. 128, p. 106393, 2020.

[7] B. Littlewood, K. Salako, L. Strigini, and X. Zhao, "On reliability assessment when a software-based system is replaced by a thought-to-be-better one," *Reliability Engineering & System Safety*, vol. 197, p. 106752, 2020.

[8] European Committee for Electrotechnical Standardization, "EN 50126: railway applications – the specification and demonstration of reliability, availability, maintainability and safety (rams)," 2017.

[9] P. Bishop, R. Bloomfield, B. Littlewood, A. Povyakalo, and D. Wright, "Toward a formalism for conservative claims about the dependability of software-based systems," *IEEE Transactions on Software Engineering*, vol. 37, no. 5, pp. 708–717, 2011.

[10] L. Strigini and A. Povyakalo, "Software fault-freeness and reliability predictions," in *Computer Safety, Reliability, and Security*, ser. LNCS, F. Bitsch, J. Guiochet, and M. Kaâniche, Eds., vol. 8153. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 106–117.

[11] X. Zhao, V. Robu, D. Flynn, K. Salako, and L. Strigini, "Assessing the Safety and Reliability of Autonomous Vehicles from Road Testing," in *the 30th Int. Symp. on Software Reliability Engineering*. Berlin, Germany: IEEE, 2019, pp. 13–23.

[12] C. Atwood, J. LaChance, H. Martz, D. Anderson, M. Englehardt, D. Whitehead, and T. Wheeler, "Handbook of parameter estimation for probabilistic risk assessment," U.S. Nuclear Regulatory Commission, Washington, DC, Report NUREG/CR-6823, 2003.

[13] B. Littlewood and L. Strigini, ""validation of ultra-high dependability..." – 20 years on," *Safety Systems, The Safety-Critical Systems Club Newsletter*, vol. 20, no. 3, May 2011. [Online]. Available: https://openaccess.city.ac.uk/id/eprint/6552/

[14] R. Soyer, "Software reliability," *WIREs Computational Statistics*, vol. 3, no. 3, pp. 269–281, 2011.

[15] B. Littlewood, "How to measure software reliability and how not to," *IEEE Transactions on Reliability*, vol. R-28, no. 2, pp. 103–110, 1979.

[16] B. Littlewood and J. Rushby, "Reasoning about the reliability of diverse two-channel systems in which one channel is 'possibly perfect'," *IEEE Tran. on Software Engineering*, vol. 38, no. 5, pp. 1178–1194, 2012.

[17] P. Popov, "Bayesian reliability assessment of legacy safety-critical systems upgraded with fault-tolerant off-the-shelf software," *Reliability Engineering & System Safety*, vol. 117, pp. 98 – 113, 2013.

[18] K. Salako, "Loss-size and reliability trade-offs amongst diverse redundant binary classifiers," in *Quantitative Evaluation of Systems*, M. Gribaudo, D. N. Jansen, and A. Remke, Eds. Springer International Publishing, 2020, pp. 96–114. [Online]. Available: https://doi.org/10.1007/978-3-030-59854-9_8

[19] J. Berger, E. Moreno, L. Pericchi, M. Bayarri, J. Bernardo, J. Cano, J. Horra, J. Martín, D. Rios, B. Betrò, A. Dasgupta, P. Gustafson, L. Wasserman, J. Kadane, C. Srinivasan, M. Lavine, A. O'Hagan, W. Polasek, C. Robert, and S. Sivaganesan, "An overview of robust bayesian analysis," *Test*, vol. 3, pp. 5–124, 06 1994.

[20] D. Insua and F. Ruggeri, *Robust Bayesian Analysis*, ser. Lecture Notes in Statistics. Springer New York, 2012. [Online]. Available: https://doi.org/10.1007/978-1-4612-1306-2

[21] X. Zhao, B. Littlewood, A. Povyakalo, and D. Wright, "Conservative claims about the probability of perfection of software-based systems," in *26th Int. Symp. on Software Reliability Eng.* IEEE, 2015, pp. 130–140.

[22] X. Zhao, B. Littlewood, A. Povyakalo, L. Strigini, and D. Wright, "Modeling the probability of failure on demand (pfd) of a 1-out-of-2 system in which one channel is "quasi-perfect"," *Reliability Engineering & System Safety*, vol. 158, pp. 230–245, 2017.

[23] P. Popov, L. Strigini, J. May, and S. Kuball, "Estimating bounds on the reliability of diverse systems," *IEEE Transactions on Software Engineering*, vol. 29, no. 4, pp. 345–359, 2003.

[24] K. Salako, L. Strigini, and X. Zhao, "Proofs of conservative confidence bounds," Tech. Rep., 2021. [Online]. Available: https://openaccess.city.ac.uk/id/eprint/25905/

[25] B. Littlewood and D. Wright, "Some conservative stopping rules for the operational testing of safety critical software," *IEEE Transactions on Software Engineering*, vol. 23, no. 11, pp. 673–683, 1997.